P&S Modern SSDs

Basics of NAND Flash-Based SSDs

Dr. Mohammad Sadrosadati
Prof. Onur Mutlu
ETH Zürich

Fall 2022

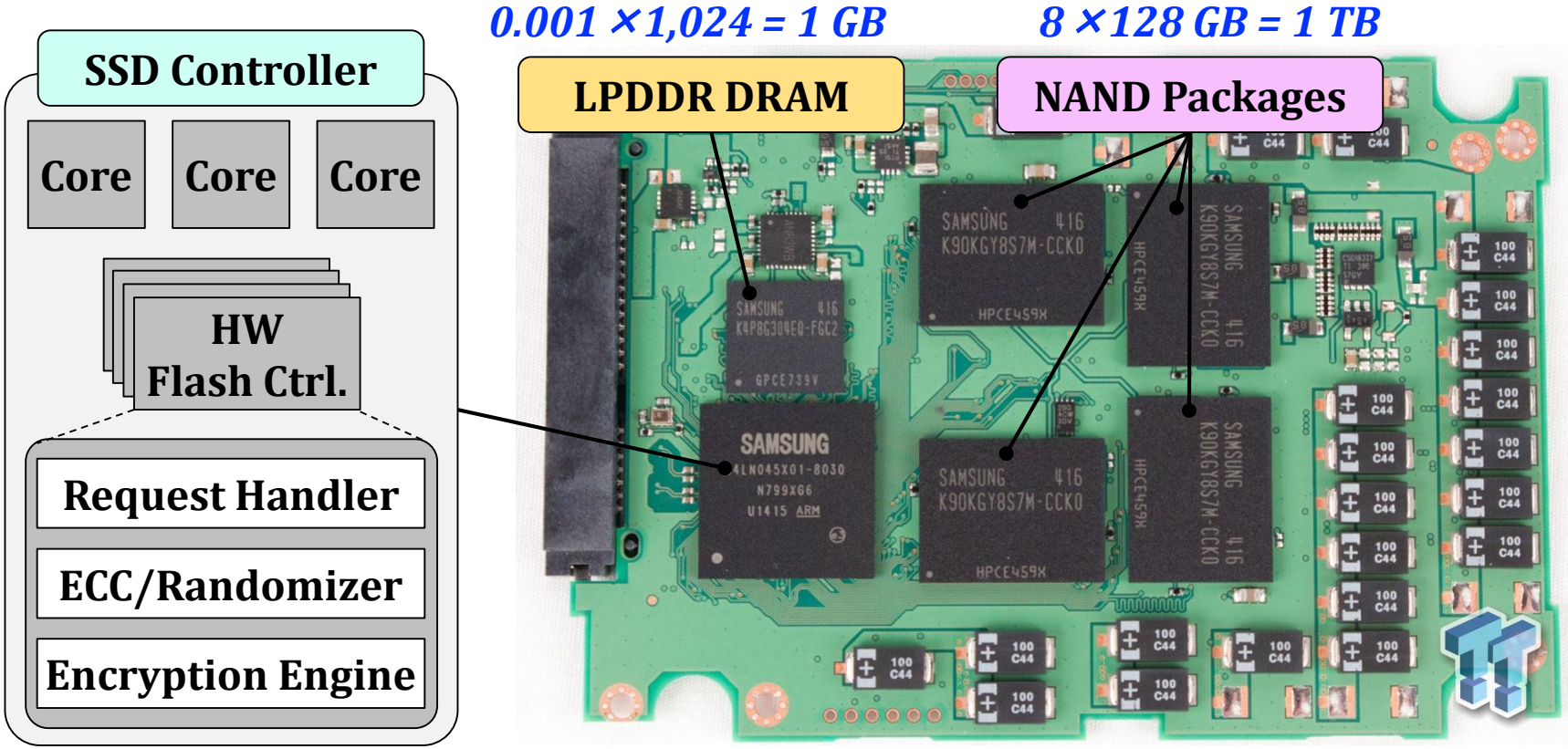
12 October 2022

Today's Agenda

- SSD Organization & Request Handling
- NAND Flash Organization

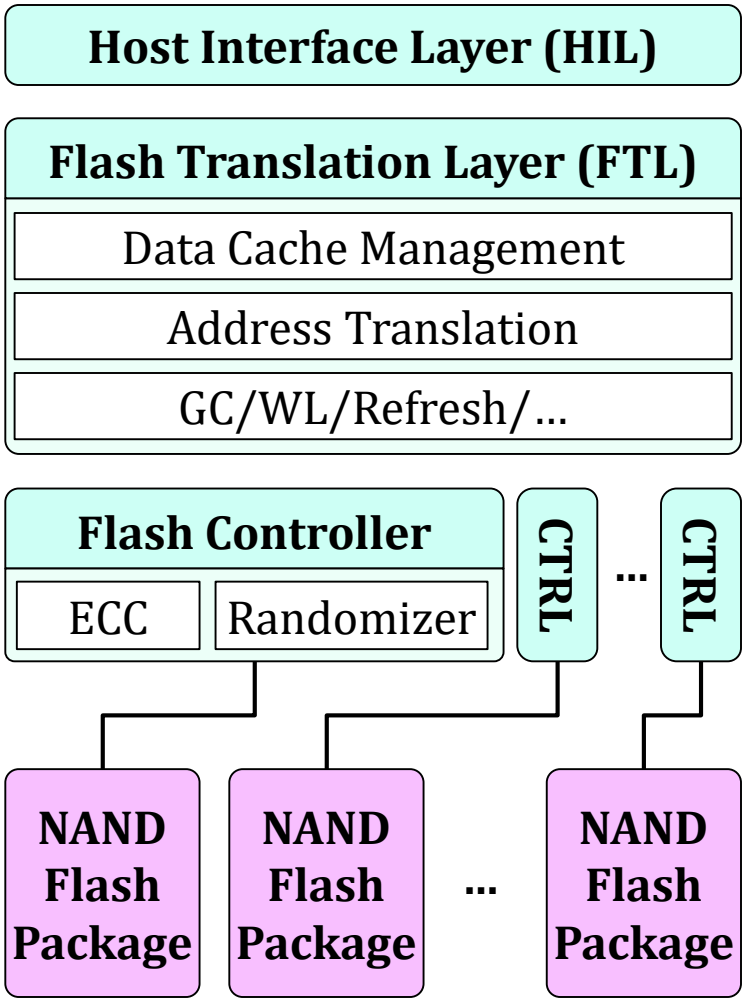
Modern SSD Architecture

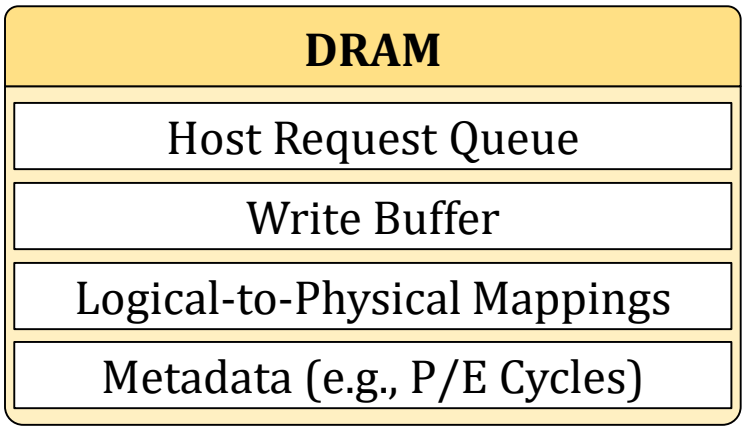
- A modern SSD is a complicated system that consists of multiple cores, HW controllers, DRAM, and NAND flash memory packages

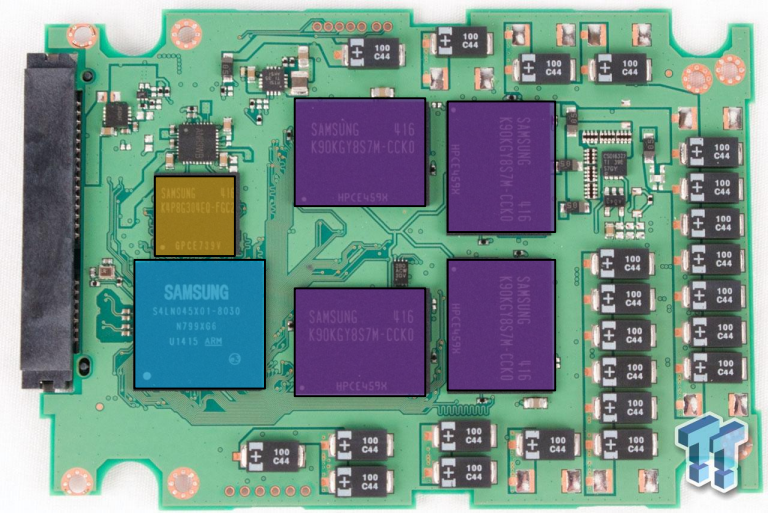


Samsung PM853T 960GB Enterprise SSD (from https://www.tweaktown.com/reviews/6695/samsung-pm853t-960gb-enterprise-ssd-review/index.html)

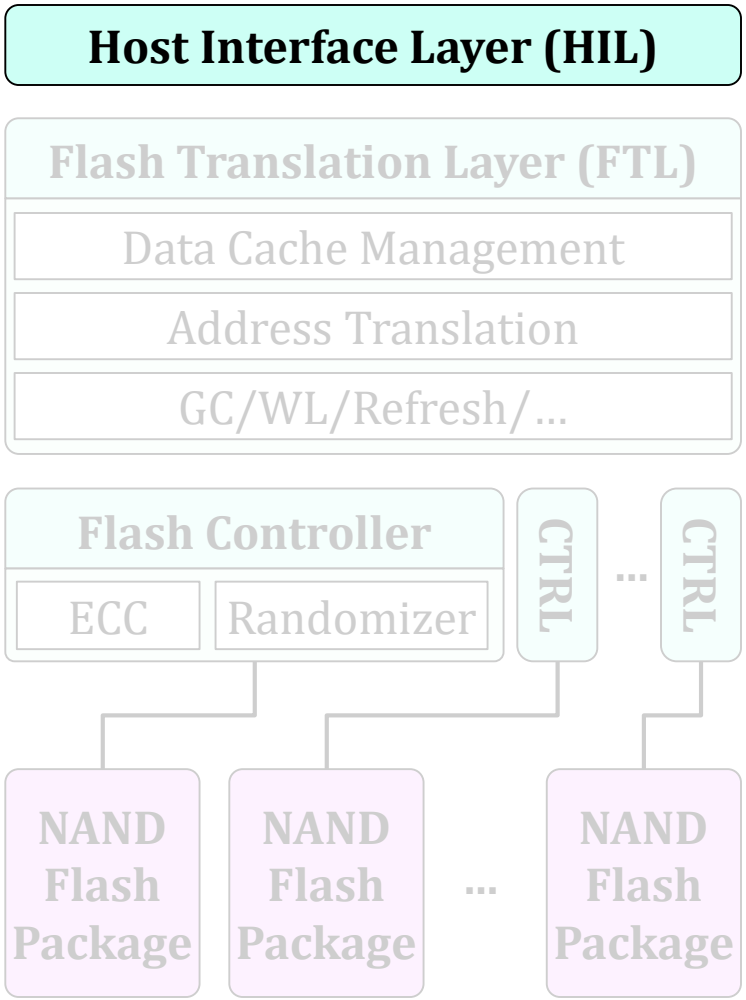
Another Overview

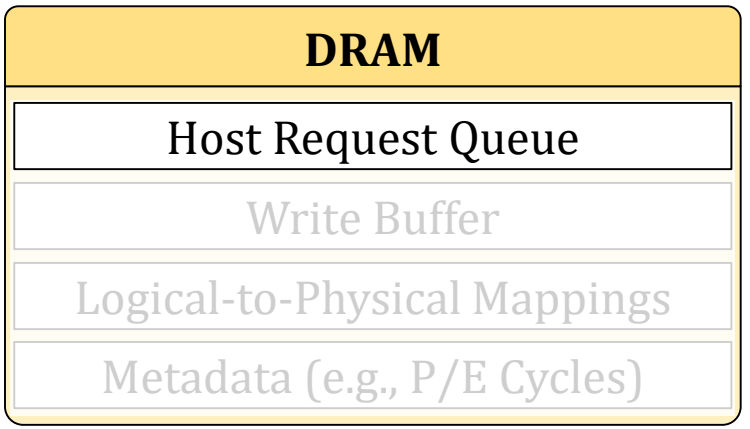






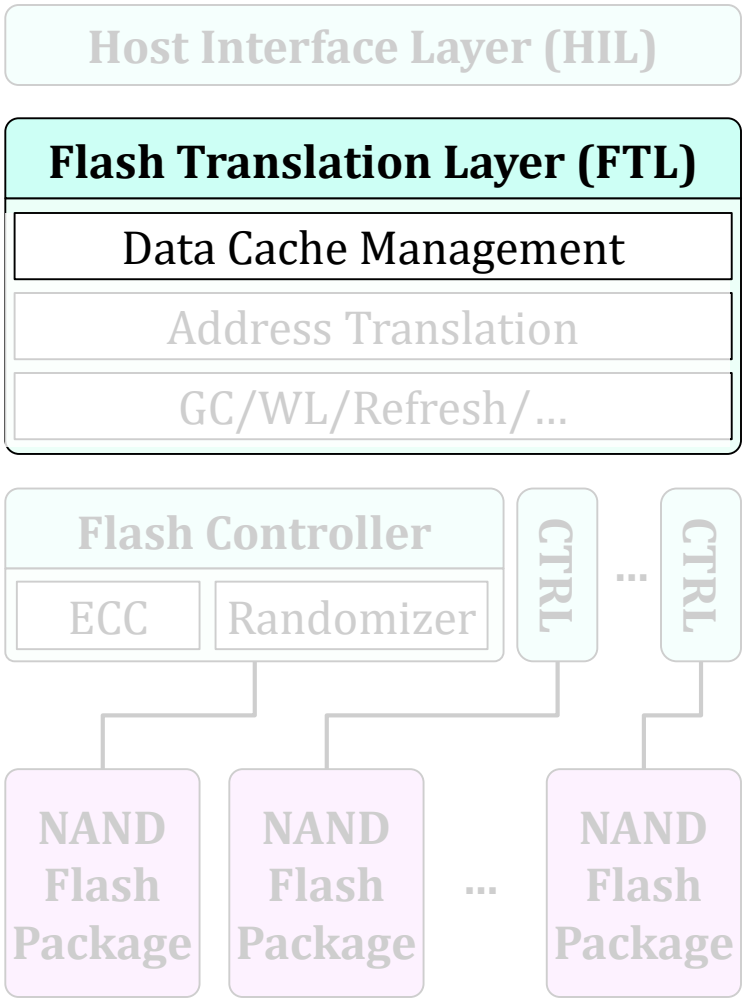
Request Handling: Write

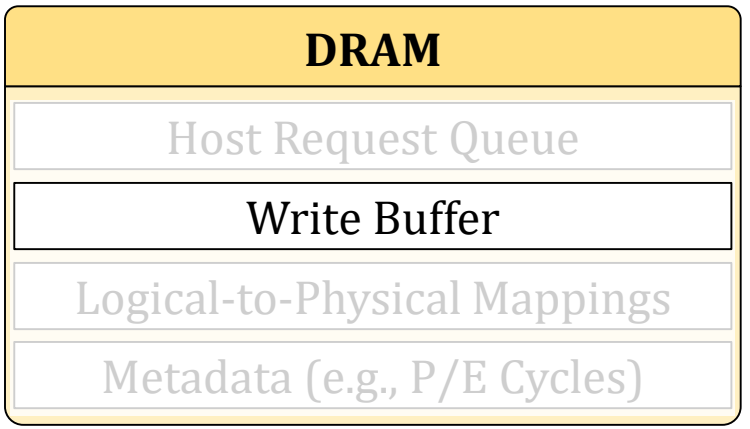




- Communication with the host operating system (receives & returns requests)
  - Via a certain interface (SATA or NVMe)
- A host I/O request includes
  - Request direction (read or write)
  - Offset (start sector address)
  - Size (number of sectors)
  - Typically aligned by 4 KiB

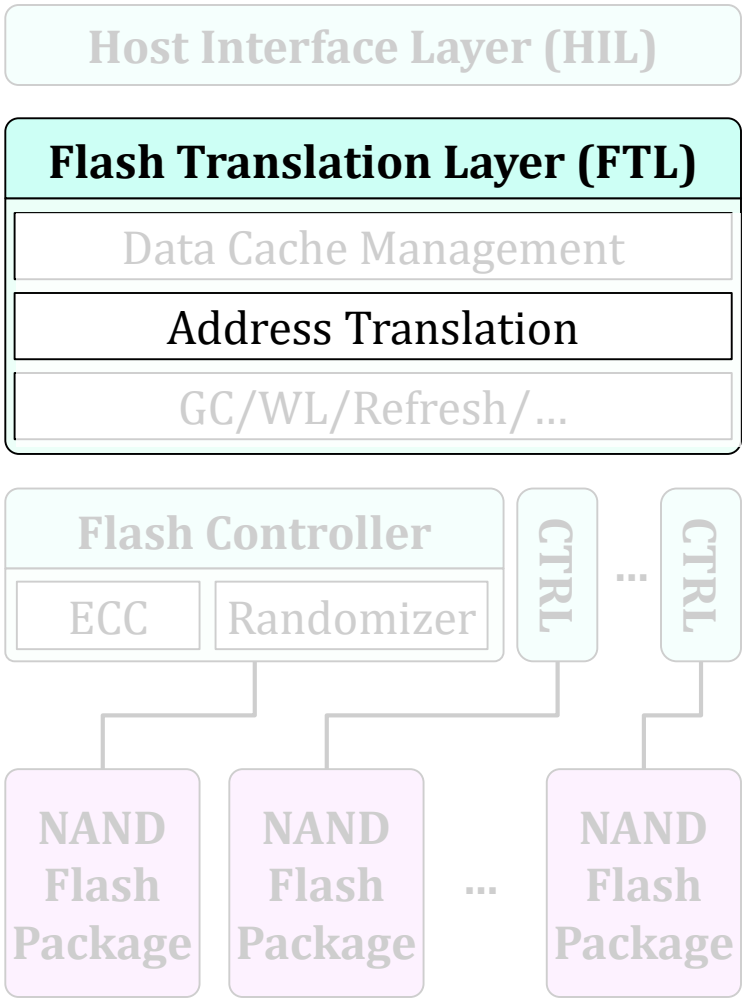
Request Handling: Write

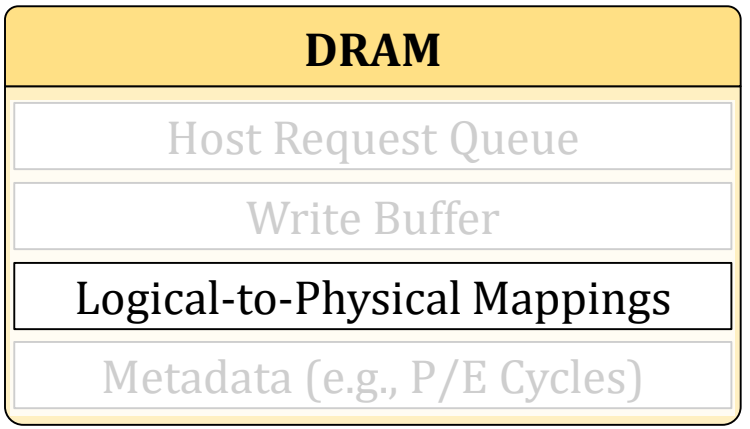




- Buffering data to write (read from NAND flash memory)
  - Essential to reducing write latency
  - Enables flexible I/O scheduling
  - Helpful for improving lifetime (not so likely)
- Limited size (e.g., tens of MBs)
  - Needs to ensure data integrity even under sudden power-off
  - Most DRAM capacity is used for L2P mappings

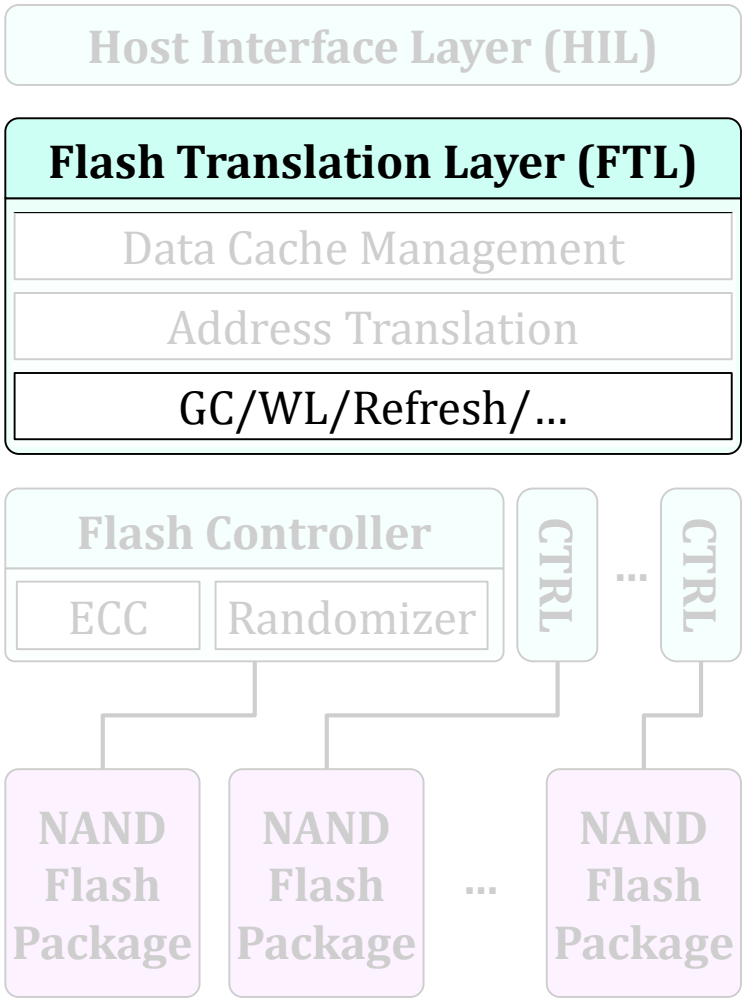
Request Handling: Write

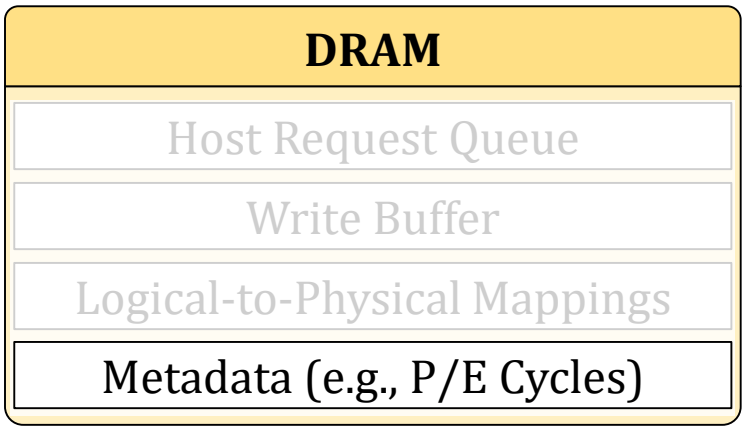




- Core functionality for out-of-place writes
  - To hide the erase-before-write property
- Needs to maintain L2P mappings
  - Logical Page Address (LPA) → Physical Page Address (PPA)
- Mapping granularity: 4 KiB
  - 4 Bytes for 4 KiB → 0.1% of SSD capacity

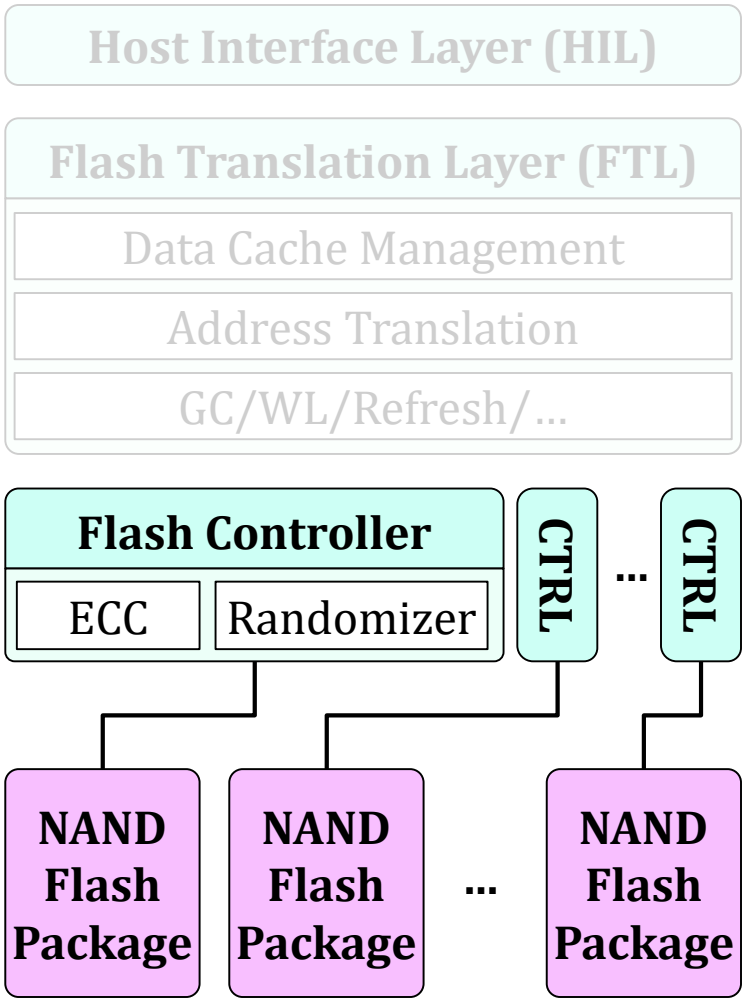
Request Handling: Write





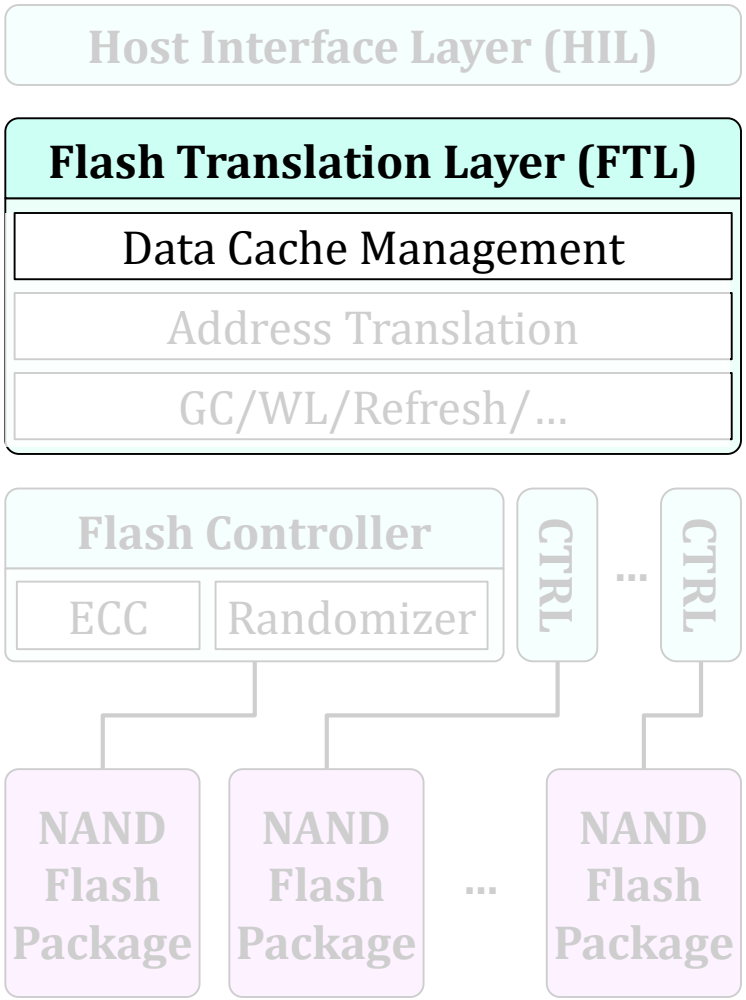
- Garbage collection (GC)
  - Reclaims free pages
  - Selects a victim block → copies all valid pages → erase the victim block
- Wear-leveling (WL)
  - Evenly distributes P/E cycles across NAND flash blocks
  - Hot/cold swapping
- Data refresh
  - Refresh pages with long retention ages

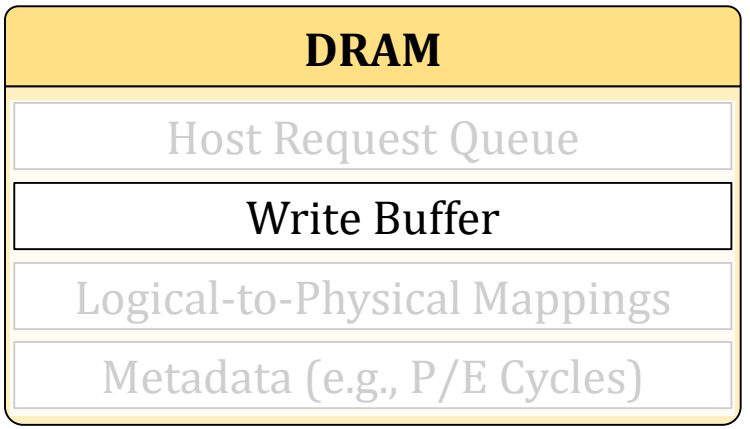
Request Handling: Write



- Randomizer
  - Scrambling data to write
  - To avoid worst-case data patterns that can lead to significant errors
- Error-correcting codes (ECC)
  - Can detect/correct errors: e.g., 72 bits/1 KiB error-correction capability
  - Stores additional parity information together with raw data
- Issues NAND flash commands

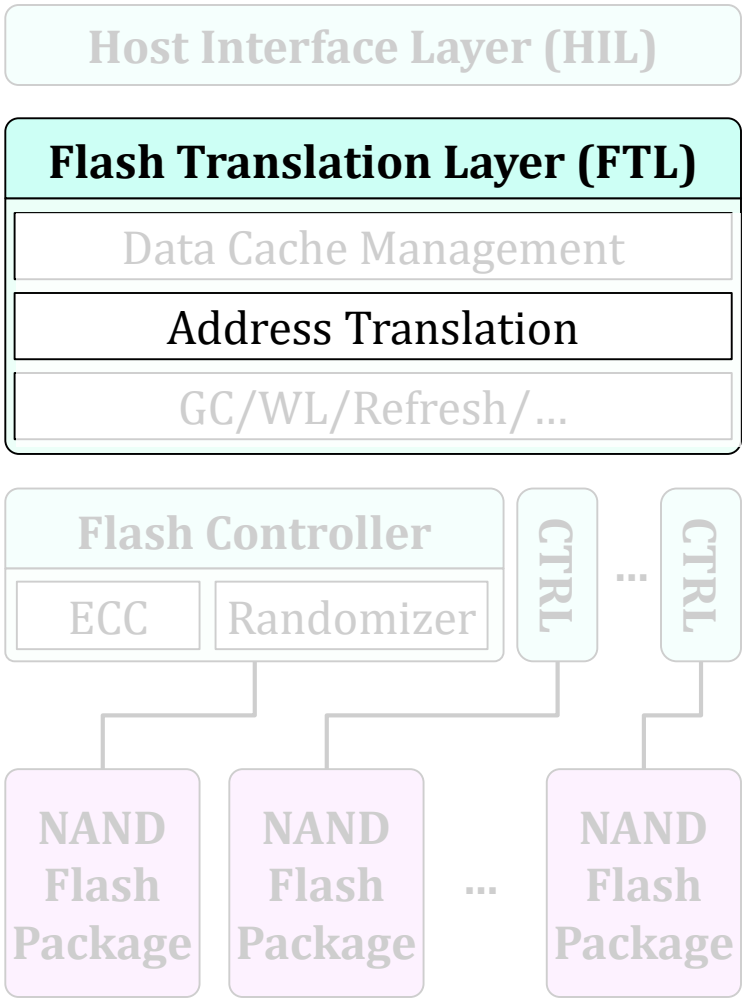
Request Handling: Read

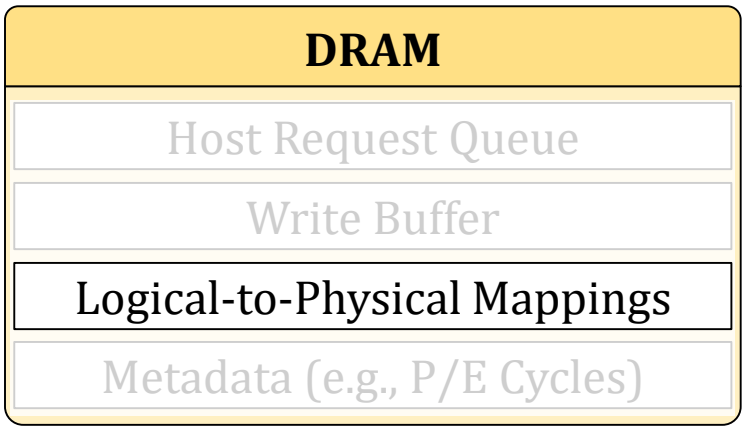




- First checks if the request data exists in the write buffer
  - If so, returns the corresponding request immediately with the data
- A host read request can be involved with several pages
  - Such a request can be returned only after all the requested data is ready

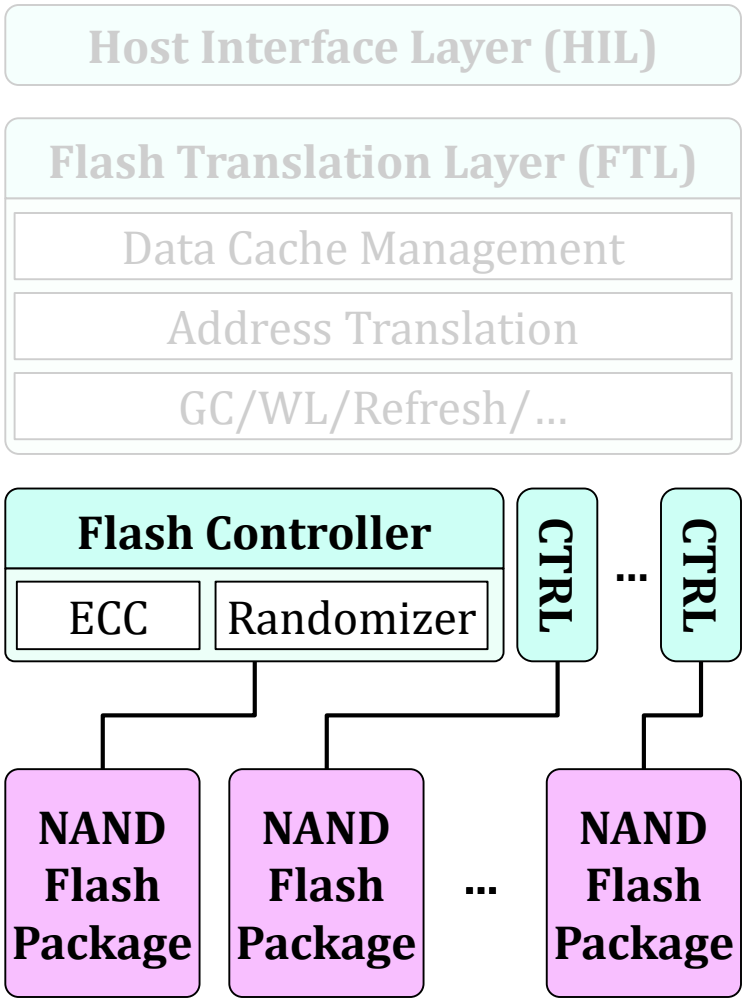
Request Handling: Read





- Finds the PPA where the request data is stored from the L2P mapping table

Request Handling: Read



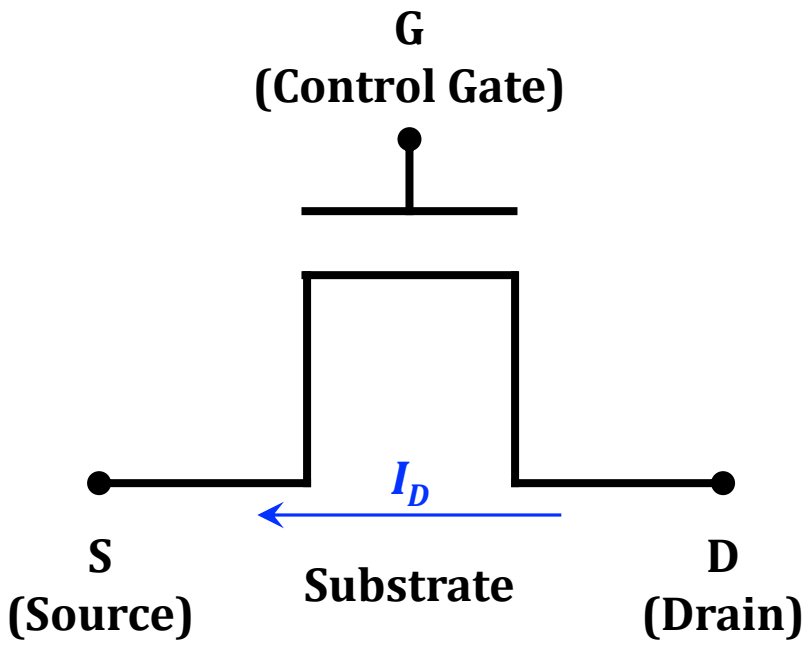
- First reads the raw data from the flash chip
- Performs ECC decoding
- Derandomizes the raw data
- ECC decoding can fail
  - Retries reading of the page w/ adjusted V_{REF}

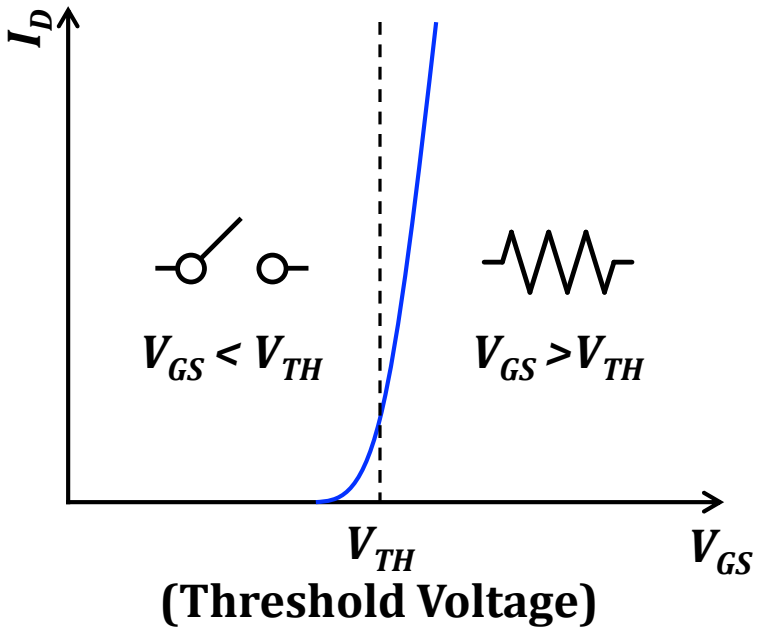
Today's Agenda

- SSD Organization & Request Handling
- NAND Flash Organization

A Flash Cell

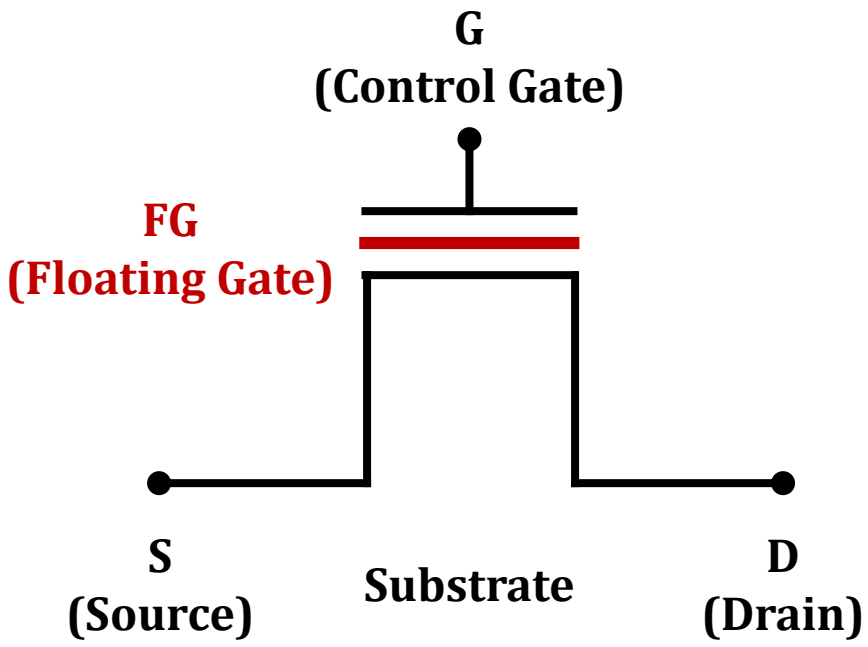
- Basically, it is a transistor

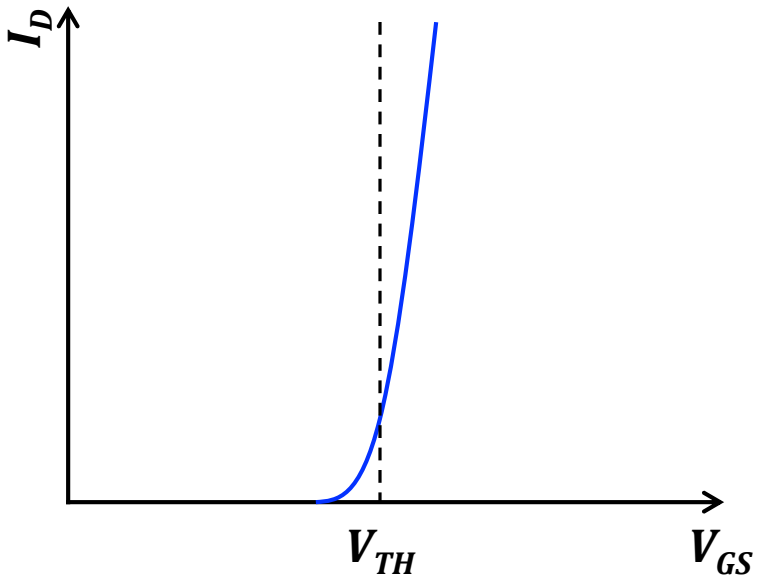




A Flash Cell

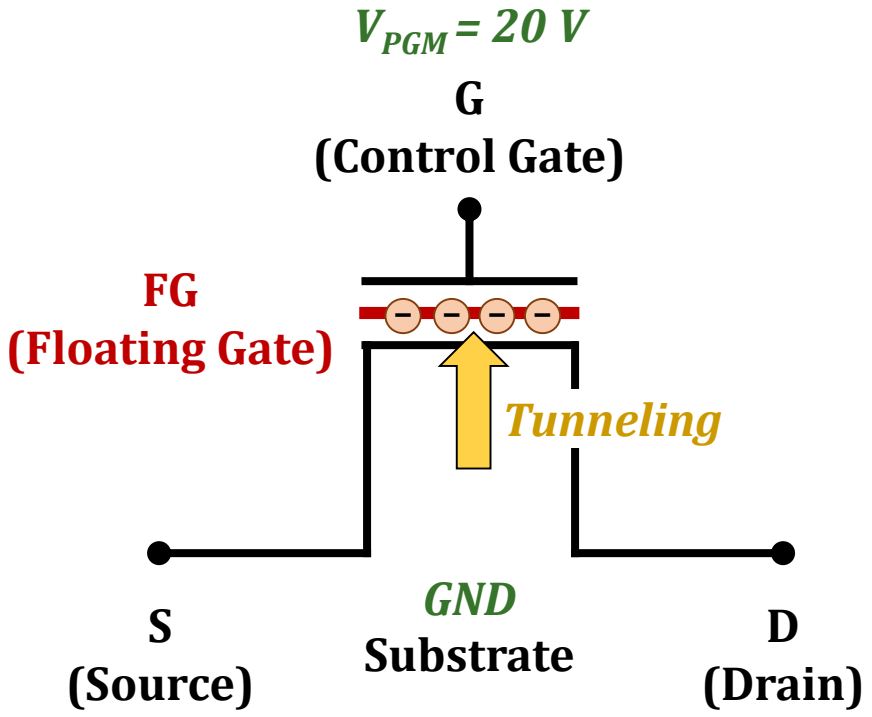
- Basically, it is a transistor
  - w/ a special material: Floating gate (2D) or Charge trap (3D)

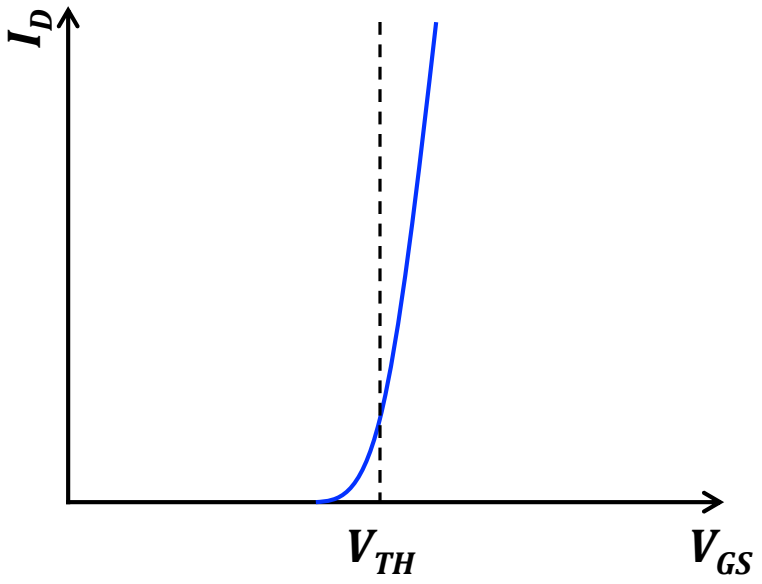




A Flash Cell

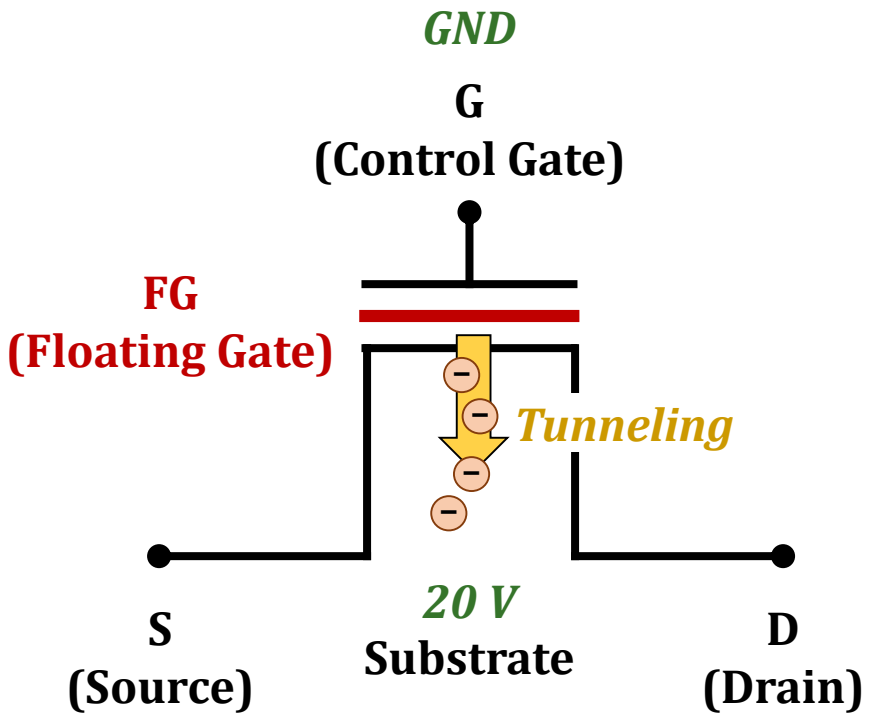
- Basically, it is a transistor
  - w/ a special material: Floating gate (2D) or Charge trap (3D)
  - Can hold electrons in a non-volatile manner

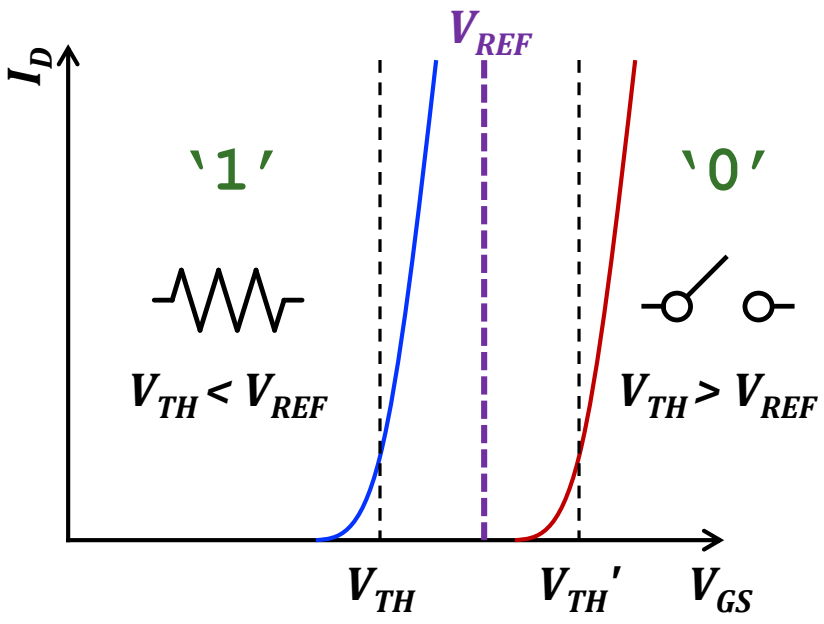




A Flash Cell

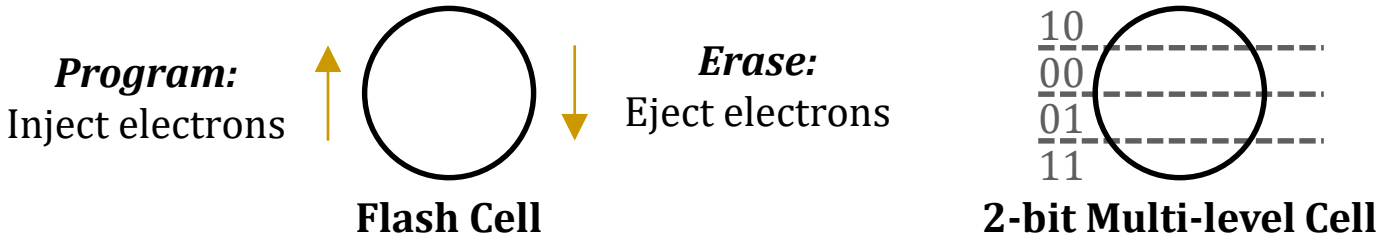
- Basically, it is a transistor
  - w/ a special material: Floating gate (2D) or Charge trap (3D)
  - Can hold electrons in a non-volatile manner
  - Changes the cell's threshold voltage (V_{TH})



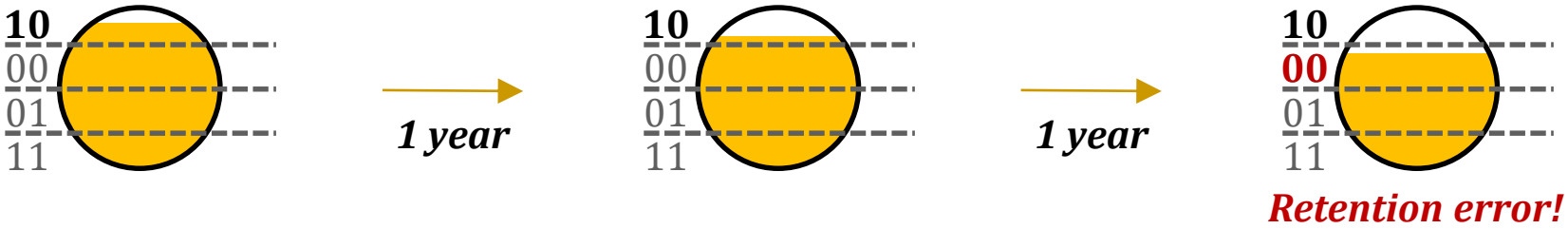


Flash Cell Characteristics

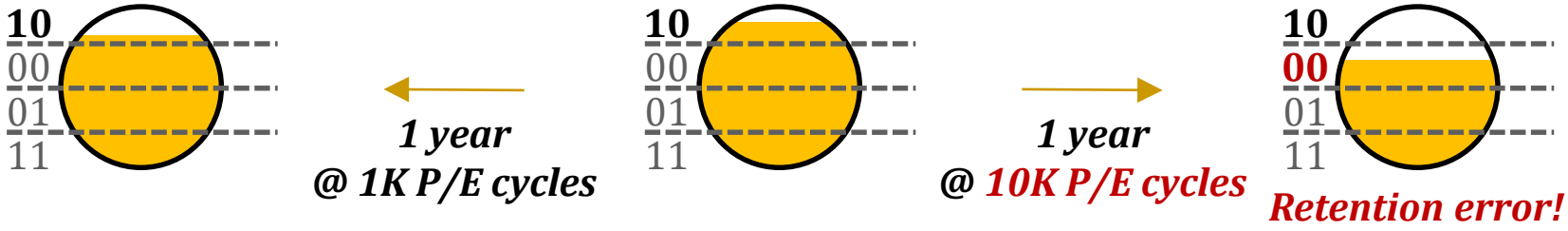
- Multi-leveling: A flash cell can store multiple bits



- Retention loss: A cell leaks electrons over time

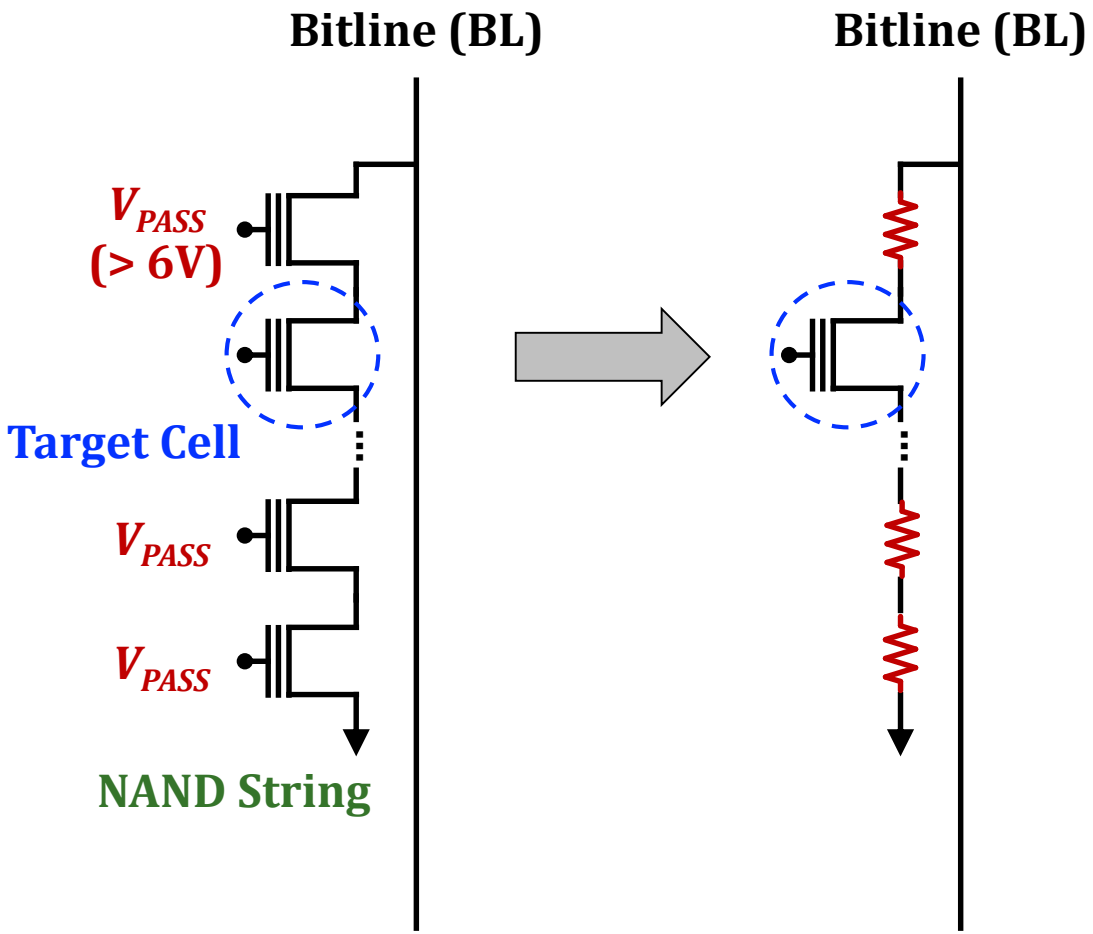


- Limited lifetime: A cell wears out after P/E cycling



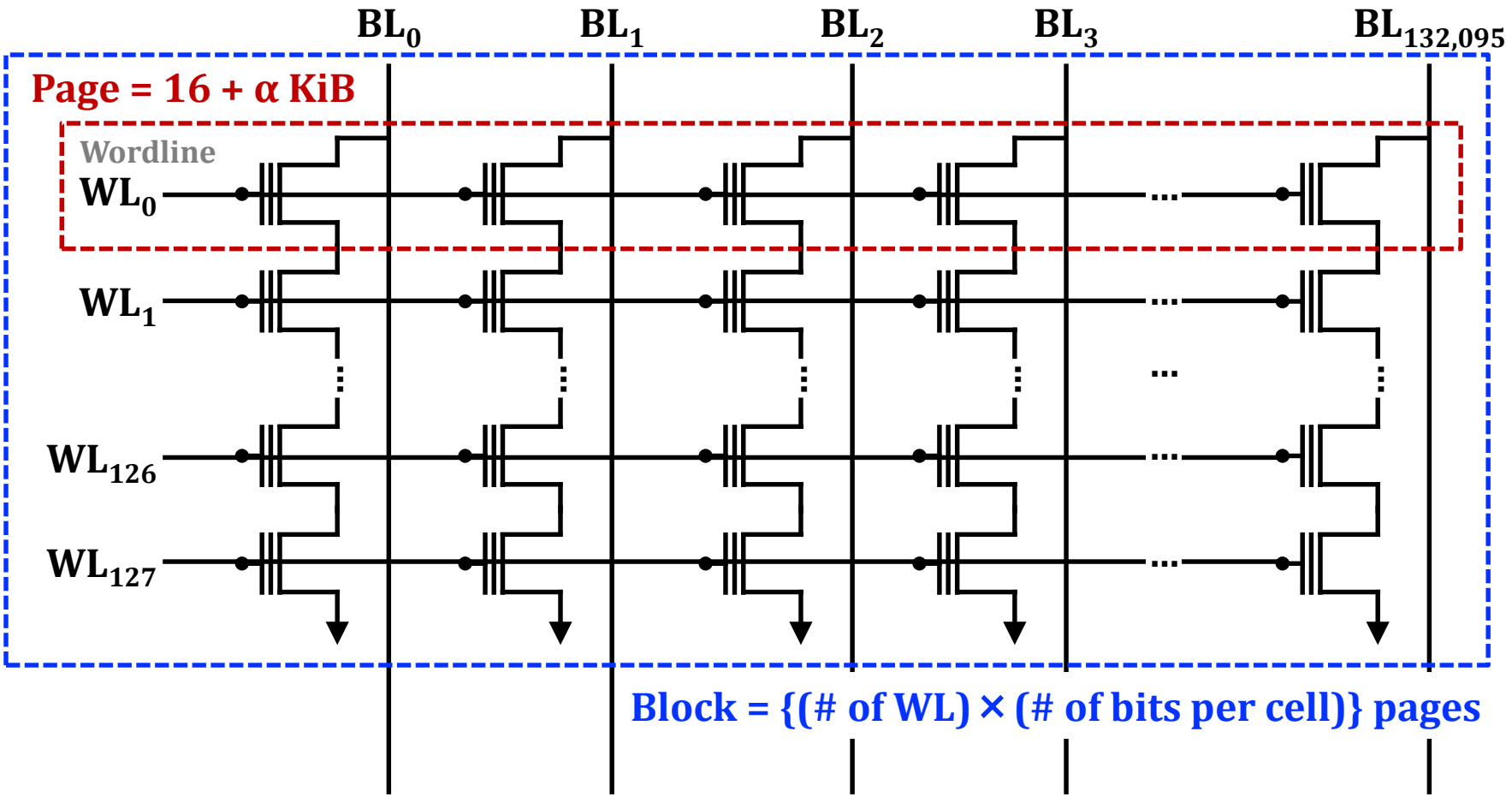
A NAND String

- Multiple (e.g., 128) flash cells are serially connected



Pages and Blocks

- A large number (> 100,000) of cells operate concurrently

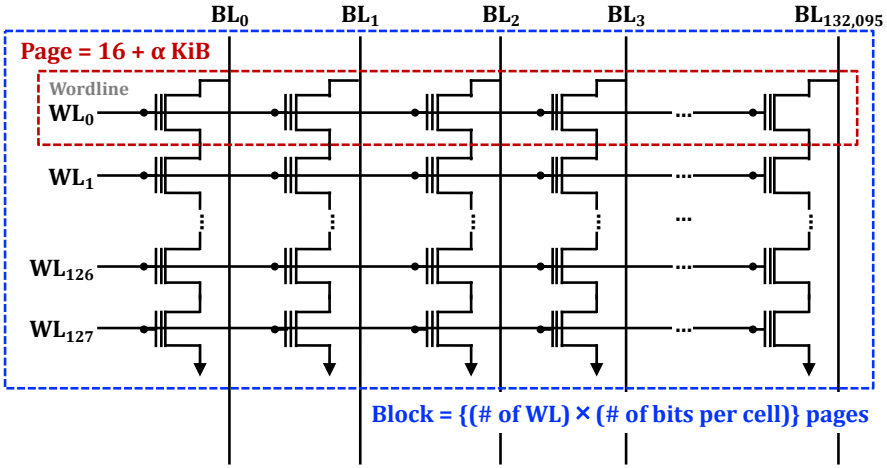


Pages and Blocks (Continued)

- Program and erase: Unidirectional
  - Programming a cell → Increasing the cell's V_{TH}
  - Eraseing a cell → Decreasing the cell's V_{TH}

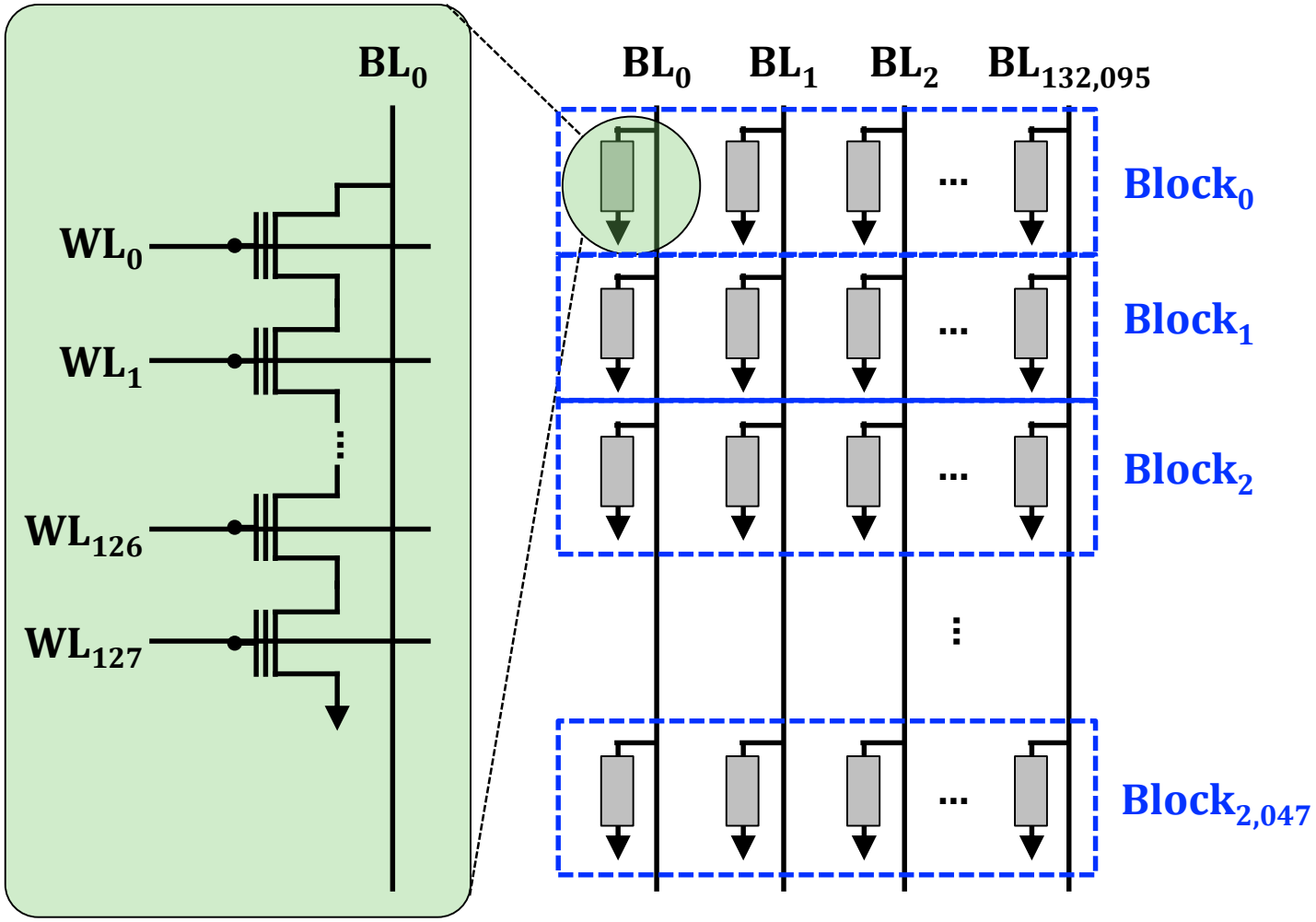
- Programming a page cannot change '0' cells to '1' cells
  → Erase-before-write property

- Erase unit: Block
  - Increase erase bandwidth
  - Makes in-place write on a page very inefficient
    → Out-of-place write & GC



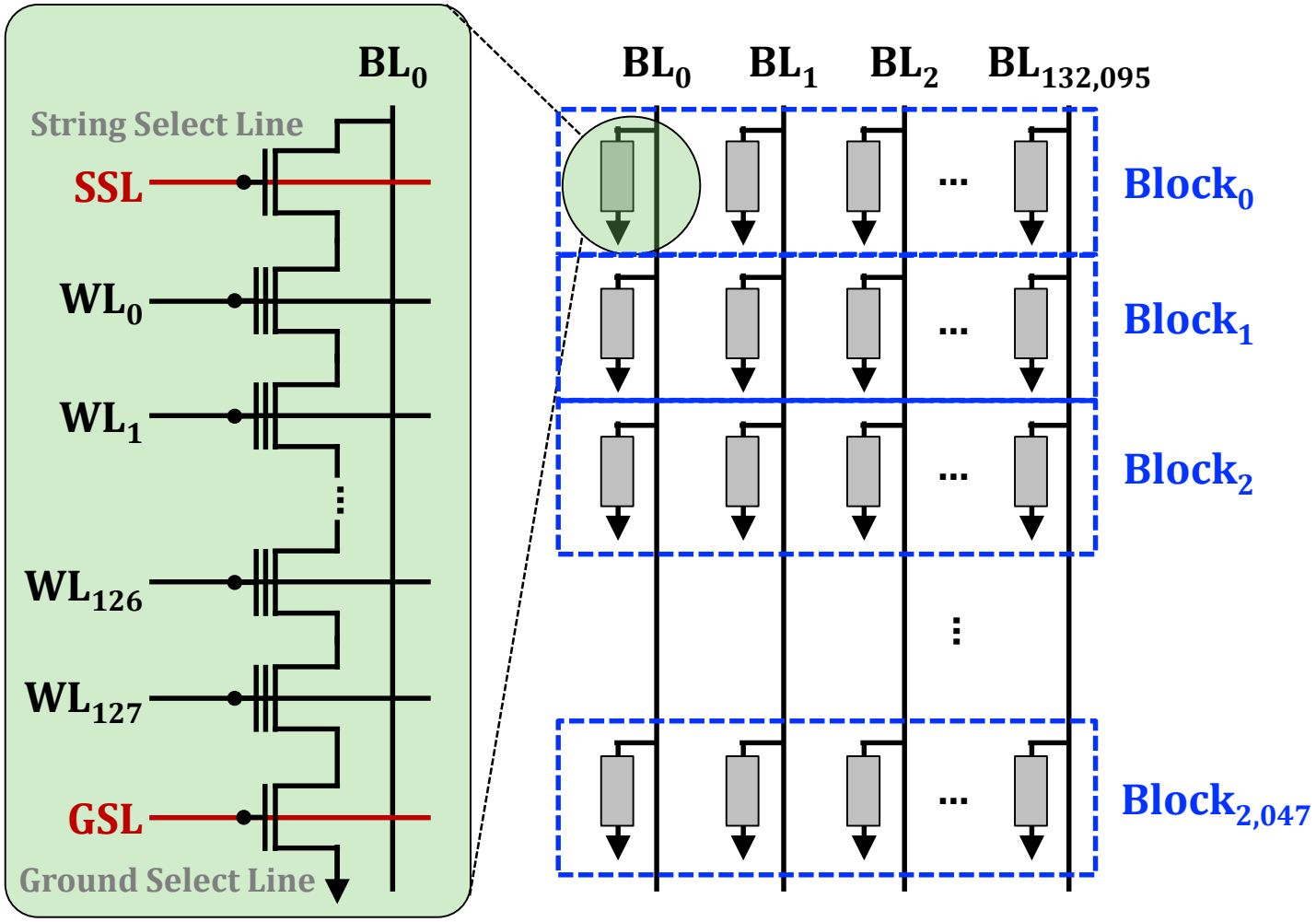
Planes

- A large number (> 1,000) of blocks share bitlines in a plane



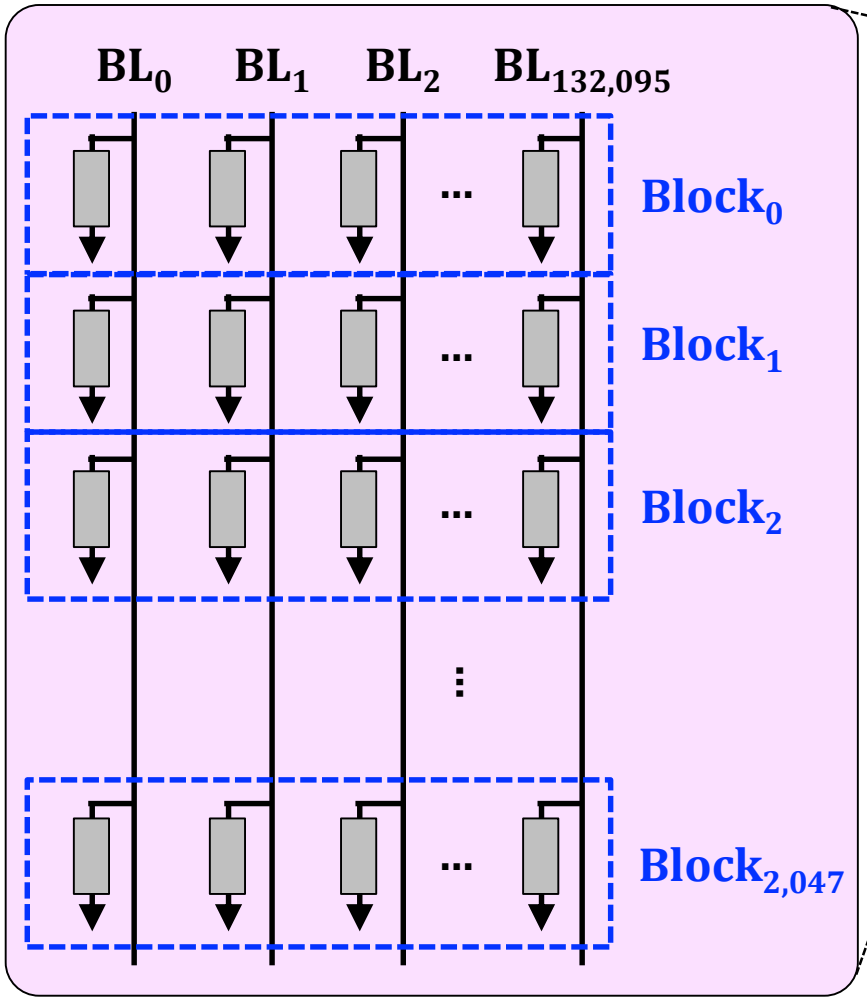
Planes

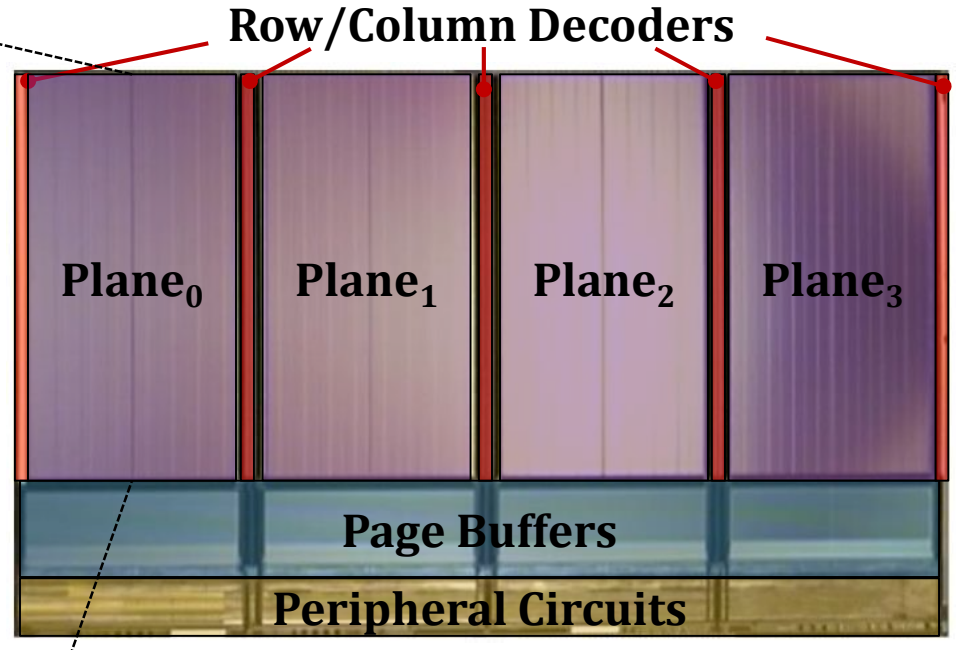
- A large number (> 1,000) of blocks share bitlines in a plane



Planes and Dies

- A die contains multiple (e.g., 2 – 4) planes





A 21-nm 2D NAND Flash Die

- Planes share decoders: limits internal parallelism (only operations @ the same WL offset)

P&S Modern SSDs

Basics of NAND Flash-Based SSDs

Dr. Mohammad Sadrosadati
Prof. Onur Mutlu
ETH Zürich

Fall 2022

12 October 2022